Computers in Biology and Medicine 59 (2015) 42–53

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

A two-step automatic sleep stage classification method with dubious range detection

Teresa Sousa*, Aniana Cruz, Sirvan Khalighi, Gabriel Pires, Urbano Nunes

Institute of Systems and Robotics (ISR-UC), Electrical and Computer Engineering Department, University of Coimbra, Portugal

ARTICLE INFO

Article history:
Received 31 July 2014
Accepted 21 January 2015

Keywords:
Automatic sleep scoring
Misclassifications detection
Subjects' variability
Dubious range
Clinical applications

ABSTRACT

Background: The limitations of the current systems of automatic sleep stage classification (ASSC) are essentially related to the similarities between epochs from different sleep stages and the subjects' variability. Several studies have already identified the situations with the highest likelihood of misclassification in sleep scoring. Here, we took advantage of such information to develop an ASSC system based on knowledge of subjects' variability of some indicators that characterize sleep stages and on the American Academy of Sleep Medicine (AASM) rules.
Methods: An ASSC system consisting of a two-step classifier is proposed. In the first step, epochs are classified using support vector machines (SVMs) spread into different nodes of a decision tree. In the post-processing step, the epochs suspected of misclassification (dubious classification) are tagged, and a new classification is suggested. Identification and correction are based on the AASM rules, and on misclassifications most commonly found/reported in automatic sleep staging. Six electroencephalographic and two electrooculographic channels were used to classify wake, non-rapid eye movement (NREM) sleep – N1, N2 and N3, and rapid eye movement (REM) sleep.
Results: The proposed system was tested in a dataset of 14 clinical polysomnographic records of subjects suspected of apnea disorders. Wake and REM epochs not falling in the dubious range, are classified with accuracy levels compatible with the requirements for clinical applications. The suggested correction assigned to the epochs that are tagged as dubious enhances the global results of all sleep stages.
Conclusions: This approach provides reliable sleep staging results for non-dubious epochs.

1. Background

The manual sleep stage classification is a labor-intensive task that involves the interpretation, by an expert, of polysomnographic (PSG) signals captured from a subject's overnight sleep session. The PSG includes electroencephalographic (EEG), electrooculographic (EOG) and electromyographic (EMG) records, respiratory effort and other physiological characteristics while a patient is asleep [1,2].

Sleep stage classification is the first step in modern diagnosis of sleep disorders. The identification of REM (rapid eye movements) sleep, NREM sleep – N1, N2 and N3 stages (non-rapid eye movement) and wake stage is performed manually by experts based on the rules of the American Academy of Sleep Medicine (AASM) reproduced in Table 1 [5,6].

Many different methods for automatic sleep staging have been proposed. ASSC algorithms consist of: data pre-processing, feature extraction and classification. The features are extracted from PSG signals and then are used as input for a classifier that provides sleep scoring. Methods for feature extraction rely mainly on frequency domain techniques, such as the spectral power of frequency bands [7]. Notwithstanding, time domain analysis [8], and more recently time-frequency domain analysis, such as wavelet [9], have been also successfully used. A wide range of machine learning techniques were already tested including linear discriminate analysis (LDA) [10], artificial neural networks (ANN) [4,11], fuzzy logic [12], decision tree classification [13], hidden Markov models (HMM) [14], clustering approaches [15] and support vector machine (SVM) [16,17]. Table 2 summarizes relevant ASSC studies and respective classification techniques. As far as we know, just a few works [7,15,17] were validated in clinical datasets. Therefore, the assessment of ASSC systems in clinical context is still very incipient.

1.1. ASSC challenges

There is no consensus about the best features and the best classification models for ASSC [11,13]. However, it is commonly accepted that the existing automatic techniques are not accurate and reliable

* Corresponding author. Tel.: +351 239 796 201; fax: +351 239 406 672.
E-mail addresses: tsousa@isr.uc.pt (T. Sousa), anianabrito@isr.uc.pt (A. Cruz), skhalighi@isr.uc.pt (S. Khalighi), gpires@isr.uc.pt (G. Pires), urbano@deec.uc.pt (U. Nunes).

http://dx.doi.org/10.1016/j.compbiomed.2015.01.017
0010-4825/

Table 1
American Academy of Sleep Medicine rules for sleep scoring.

Sleep stage	AASM rules
Wake	A. Score epochs as stage Wake when more than 50% of the epoch has alpha rhythm over the occipital region. B. Score epochs without visually discernable alpha rhythm as stage Wake if any of the following are present: (1) Eye blinks at frequency of 0.5–2 Hz; (2) reading eye movements and (3) irregular conjugate rapid eye movements associated with normal or high chin muscle tone.
N1	A. In subjects who generate alpha rhythm, score stage N1 if alpha rhythm is attenuated and replaced by low amplitude, mixed frequency activity for more than 50% of the epoch. B. In subjects who do not generate alpha rhythm, score stage N1commencing with the earliest of any of the following phenomena: (1) Activity in range of 4–7 Hz with slowing of background frequencies by ≥ 1 Hz from those of stage W; (2) vertex sharp waves; and (3) slow eye movements.
N2	A. Begin scoring stage N2 (in an absence of criteria for N3) if 1 or both of the following occur during the first half of that epoch or the last half of the previous epoch: (a) one or more K complexes unassociated with arousals and (b) one or more trains of sleep spindles. B. Continue to score epochs with low amplitude, mixed frequency EEG activity without K complexes or sleep spindles as stage N2 if they are preceded by K complexes unassociated with arousals or sleep spindles. C. End stage N2 sleep when one of the following events occurs: (1) transition to stage W; (2) an arousal (change to stage N1 until a K complex unassociated with an arousal or a sleep spindle occurs); (3) a major body movement followed by slow eye movements and low amplitude mixed frequency EEG without non-arousal associated K complexes or sleep spindles (score the epoch following the major body movements as stage N1; score the epoch as stage N2 if there are no slow eye movements); (4) transition to stage N3; and (5) transition to stage REM.
N3	A. Score stage N3 when 20% or more of an epoch consists of slow wave activity, irrespective of age.
REM	A. Score stage REM sleep in epochs with all the following phenomena: (1) low amplitude, mixed frequency EEG; (2) low chin EMG tone; and (3) rapid eye movements. B. Continue to score stage REM sleep, even in the absence of rapid eye movements, for epochs following one or more epochs of stage REM as defined in A, if the EEG continues to show low amplitude, mixed frequency activity without K complexes or sleep spindles and the chin EMG tone remains low. C. Stop scoring stage REM when one or more of the following occur: (1) there is a transition to stage Wake or N3; (2) an increase in chin EMG tone above the level of stage REM and criteria for stage N1 are met; (3) an arousal occurs followed by low amplitude, mixed frequency EEG and slow eye movements (score as stage N1; if no slow eye movements and chin EMG tone remains low, continue to score stage REM); (4) a major body movement followed by slow eye movements and low amplitude mixed frequency EEG without non-arousal associated K complexes or sleep spindles (score the epoch following the major body movement as stage N1; if no slow eye movements and the EMG tone remains low, continue to score as stage REM); and (5) one or more non-arousal associated K complexes or sleep spindles are present in the first half of the epoch in the absence of rapid eye movements(score as stage N2).

Table 2
Relevant ASSC studies. MLP: multilayer perceptron, MT: movement time, PS: paradoxical sleep, SAS: sleep apnea syndrome, SWS: slow wave sleep.

Text reference	Tested dataset	Subjects age (average)	Classification method	Classified sleep stages	Agreement *(%)
[4]	13 healthy subjects	33	MLP	Wake, N1, N2, N3, REM, MT	34; 43; 51; 82; 82; 13
[7]	4 with different pathologies from a total of 12	42.3	Clustering	Wake, N1, N2, N3, N4, REM	84; 39; 24; 81; 93; 73
[8]	47 healthy subjects	33	MLP	Wake, N1, N2, N3, PS	85; 65; 86; 93; 73
[9]	32 PSG records from MIT-BIH database	–	Regression trees	Wake, N1, N2, N3, N4, REM	93;46; 76; 58;86;77
[10]	10 healthy subjects	–	LDA	Wake, N1, N2, N3, REM	61 (ambiguous sleep), 90 (epochs with high agreement between experts)
[11]	8 healthy subjects	28	ANN	Wake, N1, N2, N3, N4, REM	84; 31; 90; 29; 77; 82
[12]	4 healthy subjects	27.5	Genetic Fuzzy	Wake, shallow sleep, deep sleep, REM	86; 84; 84; 86
[13]	–	41.8	K-NN and Decision tree	Wake, N1, N2, N3, REM, MT	80; 7; 89; 65; 82
[14]	PhysioNetdatabase	28	Hidden Markov Models	Wake, N1, N2, N3, N4, REM	51; 5; 69; 64; 92; 86
[15]	15 subjects with different severity of SAS	42	Clustering	Wake, N1, N2, N3	60 (average)
[17]	28 subjects suspected of SAS	45.5	An ensemble of five binary SVM classifiers	Wake, N1, N2, SWS, REM	96; 96; 94; 96; 95
[20]	8 healthy subejcts	28	Recurrent neural classifier	Wake, N1, N2, SWS, REM	71; 37; 97; 90; 90
[22]	20 healthy subjects	52.5	Quadratic discriminant analysis	Wake, N1, N2, SWS, REM	86; 61; 75; 93; 86
[23]	8 healthy subjects	28	Multiclass least squares SVM	Wake, N1, N2, SWS, REM	88; 76; 97; 92; 93
[25]	20 healthy subjects	21.2	Heuristic Rules	Wake, N1, N2, N3, REM	88; 35; 87; 91; 91

* The performance values here presented are the measures described in the articles. The statistical performance measures used varies from study to study.

enough to be routinely used [11]. State-of-the-art shows that sleep epochs of healthy subjects, free of sleep stage transitions and without scoring disagreement between experts, can be automatically classified using a low number of features. However, in ambiguous epochs (e.g., epochs with sleep-stage transitions) the agreement level between the results of ASSC systems and human experts is only around 60% [10].

The comparison between the studies involving patients and studies involving healthy subjects show that PSG records of patients are usually more contaminated with artifacts, have a

higher transition rate between sleep stages, and the sleep variability between subjects is higher [18]. Our previous study [19] showed that the AASC performance was around 15% lower with subjects suspected of having sleep apnea syndrome, than with healthy subjects. We concluded that the large number of movement artifacts and repetitive arousals are the main reason of performance drop. Most of the ASSC studies have improved the ASSC accuracy focusing on feature extraction [9,13,20], feature selection [8,21,22] and classification [11,12,16,17,22–24], as well on the application of heuristic rules [4,25] to solve pattern's ambiguities. In our previous work [26], a thorough analysis was made to identify the best combination of feature selection and feature extraction methods. However, neither the subject variability nor the epoch's ambiguity problems have been solved yet. As shown in Table 2, two of the three clinical studies have a classification performance too poor for clinical use [7,15]. For instance, in [15] the REM stage is not classified and in [7] the classification performance of sleep stage N1 is very low (around 38.6%). One noticeable exception is presented by Koley [17]. This system presents a classification performance higher than 94% for all sleep stages. However, in most of the sleep stages the sensitivity values remain lower than 90%, thus the sensitivity can still be improved.

1.2. Proposed system and contributions

As shown in Fig. 1, the proposed system comprises two main parts: first, the pattern recognition standard phases (pre-processing, feature extraction, training and classification, here referred as ASSCi), and second, a post-processing step based on heuristic rules to detect misclassifications. Epochs with high likelihood of misclassification are marked/tagged and a suggestion of correction is presented to the expert (Fig. 2). Identification and correction of these epochs, hereinafter called dubious epochs, are based on the AASM rules (Table 1) [5] and on the knowledge about the problems most commonly found in ASSC [2,19]. The standard sleep patterns and classification rules are used as a base to the post-processing steps since we search for a solution as generic as possible that can be applied on all patients suspected of sleep disorders.

Six EEG channels (F3, C3, O1, F4, C4, O2) and two EOG channels, right and left (ROC, LOC), are used to classify five stages (Wake, NREM sleep (N1, N2, N3) and REM sleep). The signals are classified in two steps: first with SVMs, based on a binary decision tree structure, then in the post-processing where the classification is refined. The post-processing step includes an automatic dubious range detection (DRD) and an automatic dubious range correction (DRC) module having as input the epochs classified as dubious. Detailed information is presented in Section 3.

The binary decision tree structure of the classification system yields the possibility of adjusting the features and classification parameters according to the characteristics of the different sleep stages. The system returns two outputs: epochs with likelihood of having been well classified by the SVM and epochs with a likelihood of misclassification (dubious epochs) to which a new classification is suggested (Fig. 1). Therefore, the system provides clear and reliable information to the clinical expert, classifying the major part of one sleep night data with high levels of accuracy, and tagging epochs which need a clinical expert verification.

2. Data

2.1. Participants and clinical data

The Sleep Medicine Centre of Coimbra University Hospital Centre (CHUC) provided data from all-night PSG records, which were segmented into epochs of 30s length, and scored manually by clinical experts through visual inspection. The PSG records are from fourteen subjects (10 males and 4 females) with mean age of 55.7 years (range 22–79 years, STD = 16.7 years). All subjects were diagnosed with clinical problems affecting the sleep quality [27]: obstrutive sleep apnea (8 subjects), affective disorders (2 subject), parasomnia (1 subject), periodic limb movements of sleep (1 subject), upper airway resistance syndrome (1 subject) and REM behavior disorder (1 subject). Sleep stages N2 and N3 are generally the most predominant and N1 and REM stages are the least predominant.

2.2. Data recording

PSG records were acquired by SomnoStar Pro (ViasysSensorMedics), each with duration around 8 h. The six EEG channels (F3-A2, C3-A2, O1-A2, F4-A1, C4-A1, O2-A1) and two EOG channels (ROC-A1, LOC-A2) were recorded at a sampling frequency of 200 Hz. The International 10–20 standard electrode placement system was used for EEG recording [28].

3. Methods

The proposed methodology for sleep scoring comprises two main parts that will be explained in the next sections: the pattern recognition method (pre-processing, feature extraction, training and classification) and the post-processing step based on heuristic rules to detect misclassifications (Fig. 3).

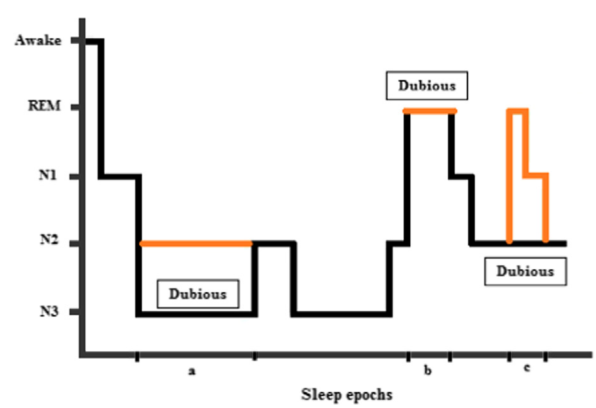

Fig. 2. Example of a final sleep hypnogram resulted of the system in Fig. 1. Dubious epochs (with the label dubious) might be assigned with a new classification, as it is illustrated in situations *a* and *c*, or the first step classification might be maintained as illustrated in situation *b*. The black hypnogram represents the initial hypnogram resulting from the ASSCi step and the parts of this hypnogram in orange color represent the classification suggested for dubious epochs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

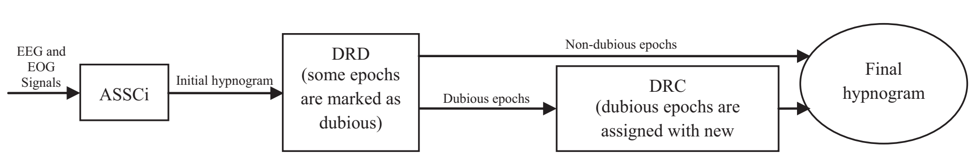

Fig. 1. Proposed system composed by ASSCi and post-processing steps. The later includes the dubious range detection (DRD) and the dubious range correction (DRC) processing modules.

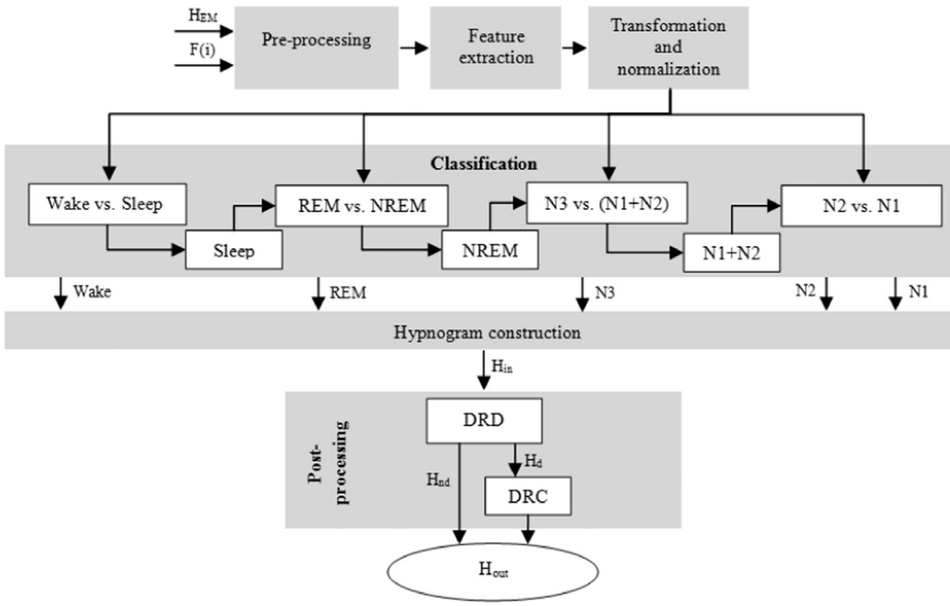

Fig. 3. Detailed structure of the proposed sleep stage classification algorithm.

3.1. Algorithm structure

The PSG signals from all channels are pre-processed, and then the features are extracted, transformed and normalized. Next, the signals are classified with an approach based on a binary decision tree structure. This approach differs from our previous multiclass SVM [26]. The decision-tree approach gives a better understanding of why and when misclassification occurs. The decision tree is organized in four decision nodes. In each node, a binary classifier is applied as illustrated in Fig. 3. The four decision nodes are wake vs. sleep; REM vs. NREM; N3 vs. (N1+N2); N1 vs. N2. The classification is performed through sequential steps. In each node, the classifier takes features and classification models that were tuned specifically for that node. The post-processing step includes the DRD and the DRC modules for detection and correction of the dubious epochs, respectively.

3.2. Dataset and resulting subsets

Variables, datasets and subset used in the ASSC algorithm described in Fig. 3 are formally presented below.

3.3. Pre-processing

The pre-processing removes the baseline drift and eliminates any linear trends [29]. Signals are filtered by a notch filter at 50 Hz and by a band-pass Butterworth filter with cutoff frequencies at 0.5 Hz and 45 Hz.

3.4. Feature extraction and selection

Feature extraction is applied to each channel using methods in frequency domain, time domain and time-frequency domain. The extracted features from EOG and from EEG are the same. The feature vector comprises 40 features extracted from each channel: 20 from maximal overlap discrete wavelet transform (MODWT), 5 from relative power, and 15 from harmonic parameters (for more details, see [21]). In each node, the best set of features were selected according to their discriminative power. The selection was made offline according to a forward/backward algorithm, aiming to reach the best classification performance. These features are discussed in Section 4.

- *Pre-processed dataset*
  $F=\{F(i,j)\ :\ i\in NP=\{1,\ldots,np_{max}\}\text{and } j\in NE=\{1,\ldots,ne_{max}\}\}$, is the set of sequential features normalized and transformed after the pre-processing step, where NP is the number of patients, np_{max} is the maximum value of NP, NE is the number of epochs, and ne_{max} is the maximum value of NE.
- *Expert hypnogram*
  $H_{EM}=\{H_{EM}(i,j)\ :\ i\in NP=\{1,\ldots,np_{max}\}\ \text{and}\ j\in NE=\{1,\ldots,ne_{max}\}\}$, is the sequential scoring of the epochs (hypnogram) made by an expert, excluding the subject that is being classified, used as model to the classification.
  The resulting subsets are as follows:
- *Initial hypnogram*
  $H_{in}=\{H_{in}(i,j)\ :\ i\in NP=\{1,\ldots,np_{max}\}\ \text{and}\ j\in NE=\{1,\ldots,ne_{max}\}\}$, is the sequential epoch classification made by the classifier (initial automatic hypnogram).
- *Dubious hypnogram*
  $H_d=\{H_d\,(i,j)\subset H_{in}\}$, is the set of epochs marked as dubious.
- *Non-dubious hypnogram*
  $H_{nd}=\{H_{in}\setminus H_d\}$, is the set of non-dubious epochs.
- *Output hypnogram*
  $H_{out}=\{DRC(H_d)\cup H_{nd}\}$, is the sequential automatic classification after the post-processing step (final automatic hypnogram).

3.4.1. Maximum overlap discrete wavelet based features

In this study, a MODWT of depth 6 with Daubechies order four (db4) is applied to every 30 s epochs with a sampling rate of 200 Hz. The frequency ranges are broken down in a decomposition of D1-D5, which correspond to δ range (<4 Hz), θ range (4–8 Hz), α range (8–13 Hz) and β range (13–30 Hz). Finally, a set of statistical MODWT-based features are extracted to represent the time-frequency distribution of the EEG, and EOG signals: energy, percentage of energy, frequency and temporal features [30].

3.4.2. Relative spectral power (RSP)

Spectral analysis provides some of the most important features. For each signal X, an FFT squared modulus estimator was applied to estimate the power spectral density (PSD). The spectrum is divided into five frequency sub-bands as described for MODWT. For each frequency sub-band, the RSP is computed. This parameter is given by the ratio between the sub-band spectral power and the total spectral power, i.e., the sum of all five BSP sub-bands [31].

3.4.3. Harmonic parameters

Harmonic Parameters of the EEG and EOG signals include three parameters: the center frequency, the bandwidth and the spectral value at center frequency for each sub-band [32].

3.5. Features transformation and normalization

To reduce the influence of extreme values, the extracted features are transformed and normalized. The log transform

$$Y = \log(X) \tag{1}$$

was adopted in the overall sleep staging, since the empirical analyses revealed that this transform produced consistently the best classification results [33]. In (1) $\boldsymbol{X}$ denotes the feature matrix, and the output is

$$Y = \{y_{ib}; i = 1, 2, \ldots, NP \text{ and } b = 1, 2, \ldots, NF\} \tag{2}$$

where NP and NF denote the number of subjects and the number of features, respectively. Each feature of the transformed matrix $\boldsymbol{Y}$ is independently normalized

$$\overline{Y}_{ib} = y_{ib}/(\max(y_i) - \min(y_i)) \tag{3}$$

where y_i is a vector of each independent feature [34].

3.6. Classification

In this study, two widely used classifiers were compared: LDA [35] and SVM [36]. The LDA is a statistical classifier, which is based on the “between class” and “within class” scatter matrices. LDA learns a linear classification boundary in the input feature space which is based on generative models. SVM is a discriminative-based classifier that usually uses kernels to construct linear classification boundaries in high-dimensional spaces. The SVM classifier constructs a hyper-plane as a decision surface such that the margin of separation between different classes is maximized.

The performance comparison between the two classifiers helps to validate the independence of the methods used for feature selection and feature extraction, in respect to the classifier. They were both tested using a binary decision tree structure (BDT), and therefore henceforth they will be called BDT-SVM and BDT-LDA.

3.7. Post-processing

The two main post-processing modules, DRD and DRC, are detailed in the next sections. The post-processing uses a set of rules to detect dubious epochs. The rules were all defined *a priori* from a deep analysis of the dataset.

3.7.1. Dubious range detection (DRD)

The post-processing rules (PPr) were defined from a systematic analysis of the following points: (1) the acceptable patterns for each sleep stage; (2) the acceptable transitions among sleep stages; and (3) the acceptable sleep stages sequences (temporal windows with groups of 3–5 consecutive epochs), according to the AASM rules. The PPr are organized in three groups:

1. Sleep stages sequence – the hypnogram H_{in} is checked to verify if there are sleep sequences impossible/unlikely to occur.
2. Sleep transitions – changes on individual and combination of features are checked to verify if they are in admissible ranges for each sleep transition.
3. Standard patterns of each sleep stage – the agreement between the features and the classification output of each epoch is verified.

Despite the considerable night-to-night variation among different individuals, certain generalizations can be made based on the standard sleep patterns [37]. Sleep normally begins with stages N1 and N2, and then may progress to deeper stage N3. REM sleep occurs in discrete episodes typically beginning about 90 min after a sleep onset and periodically at intervals of around 90 min. The REM episodes tend to lengthen over the night. Therefore, there are some sequences of sleep stages which are not possible, for example, someone awake can not go into deep sleep directly, neither in REM sleep (Fig. 4). These heuristic rules, can automatically identify classification errors made by SVM or LDA classifiers. According to the normal sleep cycles, the transitions from N1 to N3 should not occur without passing by the N2 stage, and a transition from N3 to REM is also not common. However, in subjects with sleep disorders, the arousals often lead to some unusual transitions. To improve the detection of dubious epochs, the rules consider the information of more than two sequential epochs.

Table 3 presents the main EEG patterns for each sleep stage [5]. Some patterns are common to different stages, which is a cause of misclassification. For example, sleep transition PPr verify if in transitions from wake to N1 the alpha activity decreases, and if

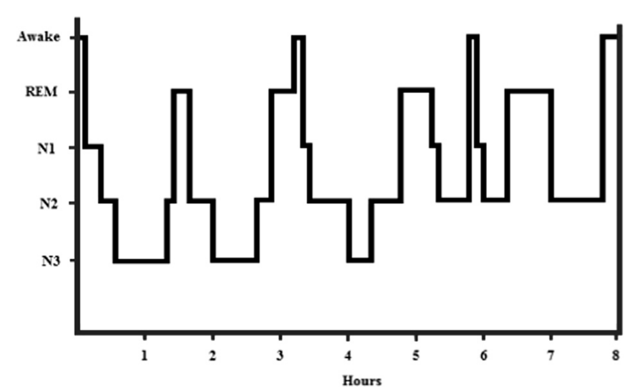

Fig. 4. Sleep cycles overnight, with deep sleep early on and more REM toward morning.

Table 3
Summary of EEG patterns for different sleep stages.

	Delta (<4 Hz)	Theta (4–7 Hz)	Alpha (8–13 Hz)	Beta (>13 Hz)	Other patterns
Wake			x	x	
N1		x	x		Vertex waves
N2		x			K complex; sleep spindles
N3	x				Sleep spindles may persist
REM		x			Sawtooh waves

theta activity increases, and in transitions from N2 to N3, if delta activity increases. The PPr based on standard patterns described in Table 4 include rules such as “if one epoch is classified as N3 sleep stage the delta activity should be its main type of activity”.

The rules are presented in Table 4 in a formal way. For example, the rule $C(j-1)=0 \cap C(j)=2$ means that if one epoch (j) was classified as N2 and the previous epoch ($j-1$) was classified as wake, this epoch (j) is tagged as dubious. The SVM classification output ($C=\{0,1,2,3,5\}$) represents the sleep stages wake, N1, N2, N3 and REM, respectively. In order to withstand the effects of subjects' variability, these rules vary according to characteristics of the brain activity of each subject, by defining a patient profile. Three patient profiles were defined (DRp1, DRp2 and DRp3), based on the amplitude of EEG delta rhythms since these rhythms have shown to be the ones that differ the most between subjects [38].

3.7.2. *Dubious range correction (DRC)*

The automatic DRC (Fig. 5) uses information of typical manual scoring of an expert (H_{EM}). The features of dubious epochs (H_d) are compared with the features of epochs scored by an expert in the

Table 4
Post-processing rules to define the dubious range epochs (j) to patient profile 1 (DRp1), patient profile 2 (DRp2) and patient profile 3 (DRp3). Where: C – Cassification result; δ – relative power of delta activity; θ – relative power of theta acctivity; α – relative power of alpha and β – relative power of beta.

	PPR		
	DRp1 $\delta < 0.7$	DRp2 $\delta > 0.8$	DRp3 $0.7 \le \delta \le 0.8$
Sleep sequence	$C(j-1)=0 \cap C(j)=2$ $C(j)=0 \cap C(j+1)=5$ $C(j-1)=1 \cap C(j)=5 \cap C(j+1)=1$ $C(j-1)=3 \cap C(j)=2 \cap C(j+1)=3$ $C(j-1)=1 \cap C(j)=3$ $C(j)=2 \cap C(j+1)=5$ $C(j-2)=3 \cap C(j-1)=3 \cap C(j)=2 \cap C(j+1) = 3 \cap C(j+2)=3$	$C(j-1)=0 \cap C(j)=2$ $C(j)=0 \cap C(j+1)=5$ $C(j-1)=1 \cap C(j)=2 \cap C(j+1)=1$ $C(j-1)=3 \cap C(j)=2 \cap C(j+1)=3$ $C(j)=2 \cap C(j+1)=5$	$C(j-1)=0 \cap C(j)=2$ $C(j)=0 \cap C(j+1)=5$ $C(j-1)=1 \cap C(j)=5 \cap C(j+1)=1$ $C(j-1)=1 \cap C(j)=2 \cap C(j+1)=1$ $C(j-1)=3 \cap C(j)=2 \cap C(j+1)=3$ $C(j)=2 \cap C(j+1)=5$ $C(j-2)=3 \cap C(j-1)=3 \cap C(j)=2 \cap C(j+1) \cap =3\ C(j+2)=3$
Sleep transitions	$C(j-1)=0 \cap C(j)=1 \cap C(j+1)=0 \cap \delta(j) \le 0.54$ $C(j-1)=5 \cap C(j)=1 \cap (\delta(j)-\delta(j-1))>0$ $C(j-1)=5 \cap C(j)=0 \cap (\delta(j)-\delta(j-1))<0$ $C(j)=5 \cap C(j+1)=1 \cap (\alpha(j)-\alpha(j+1))>0$ $C(j)=2 \cap C(j+1)=3 \cap (\beta(j)-\beta(j+1))<0$	$C(j-1)=2 \cap C(j)=3 \cap (\delta(j)-\delta(j-1))<0$ $C(j-1)=5 \cap C(j)=2 \cap (\delta(j)-\delta(j-1))<0$ $C(j)=5 \cap C(j+1)=1 \cap (\alpha(j)-\alpha(j+1))>0$ $C(j)=2 \cap C(j+1)=3 \cap (\beta(j)-\beta(j+1))<0$	$C(j-1)=0 \cap (j)=1 \cap C(j+1)=0 \cap \delta(j) \le 0.5$ $C(j-1)=5 \cap C(j)=2 \cap (\delta(j)-\delta(j-1))<0$ $C(j-1)=5 \cap C(j)=0 \cap (\delta(j)-\delta(j-1))<0$ $C(j)=2 \cap C(j+1)=3 \cap (\beta(j)-\beta(j+1))<0$
Standard patterns	$C(j)=2 \cap \beta(j)<0.06$ $C(j)=2 \cap \theta(j) \le 0.32$ $C(j)=2 \cap \delta(j) \ge 0.95 \cap \theta(i) \ge 0.55$ $C(j)=1 \cap \theta(j) \le 0.29$ $C(j)=1 \cap \alpha(j)>0.76$	$C(j)=2 \cap \theta(j)>0.38$ $C(j-1)=5 \cap C(j)=2 \cap \alpha(j) \ge 0.35$ $C(j)=2 \cap \delta(j) \ge 0.95 \cap \theta(j) \ge 0.55$ $C(j)=1 \cap \theta(j) \le 0.28$ $C(j)=3 \cap \theta(j) \ge 0.61$ $C(j)=5 \cap \beta(j) \le 0.11$	$C(j)=1 \cap \beta(j) \ge 0.8$ $C(j)=2 \cap \theta(j) \le 0.37$ $C(j)=2 \cap \beta(j)>0.21$ $C(j-1)=5 \cap C(j)=2 \cap \alpha(j) \ge 0.48$ $C(j)=2 \cap \delta(j) \ge 0.95 \cap \theta(j) \ge 0.55$ $C(j)=1 \cap \delta(j) \le 0.48$ $C(j)=2 \cap \alpha(j) \ge 0.61$ $C(j)=5 \cap \beta(j) \le 0.09$

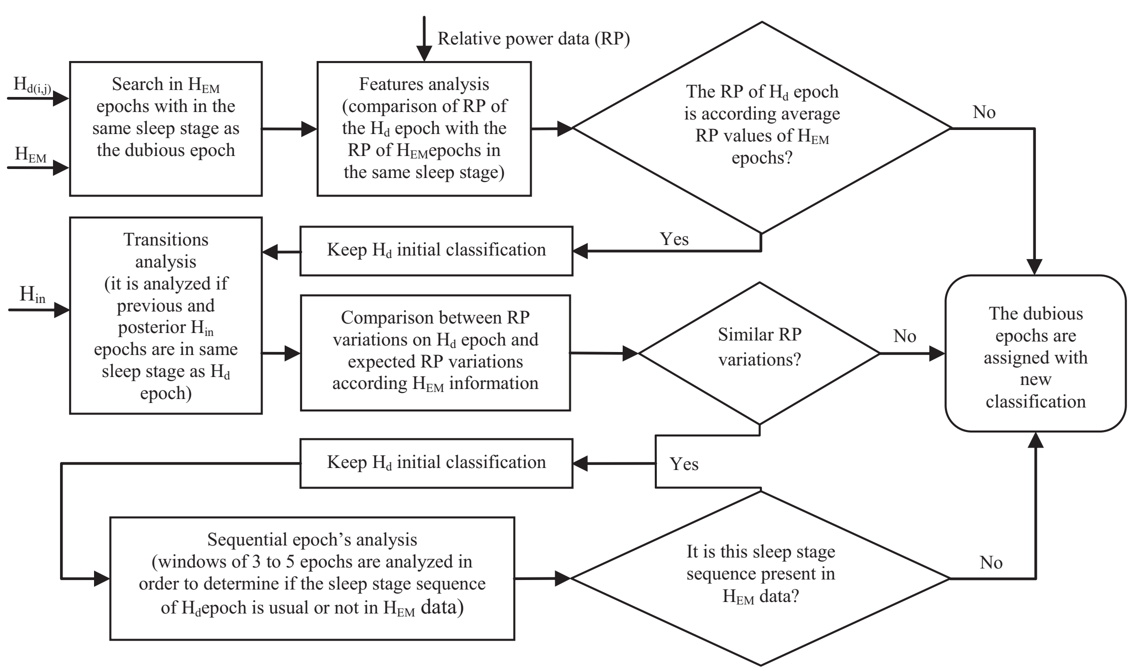

Fig. 5. DRC process.

Table 5
Results of classification accuracy for the selected combination of features (average results of all subjects).

Features	Decision node			
	Wake vs. sleep	REM vs. NREM	N3 vs. N2+N1	N2 vs. N1
Relative power	94.35	93.48	90.40	87.89
Relative power and MODWT	92.72	92.73	89.65	88.90
Relative power, MODWT and harmonic parameters	93.61	94.01	90.58	89.49
Selection:				
Relative power	94.35	–	–	–
Relative power and harmonic parameters	–	95.07	–	–
Relative power and harmonic parameters	–	–	91.71	–
Relative power, MODWT and harmonic parameters	–	–	–	89.49

same stage. If the dubious epoch is included in a sleep transition, this transition is analyzed and the agreement between sleep stages' transition and features variations is verified as follows. If the classification result is similar to a H_{EM} epoch, with the same characteristics, the suggested correction coincides to H_{in} (i.e., the hypnogram before the post-processing). On the other hand, if the dubious epoch classification differs significantly from the H_{EM} epochs, it is suggested a new classification, corresponding to the sleep stage for which the characteristics coincide the most.

3.8. Performance assessment

The performance of the algorithm was accessed using leave-one subject-out cross-validation (LOOCV) [39]. To calculate the classification performance for one subject, the algorithm uses the data of the other subjects for training. To measure the classification performance, the balanced accuracy values (considering a multi-class confusion matrix) were calculated as follows [40]:

$$\text{Balanced accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{4}$$

where

$$\text{Sensitivity} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \tag{5}$$

$$\text{Specificity} = \frac{\text{True negatives}}{\text{False positives} + \text{True negatives}}. \tag{6}$$

4. Experimental results

4.1. Feature and classifier selection

The approach was tested using a dataset with fourteen subjects as mentioned in Section 2. The EEG and EOG signals, extracted from the PSG records, were filtered, and segmented into 30 s epochs as described in the pre-processing step. Then, features in the frequency domain (relative power and harmonic parameters), time-frequency domain (MODWT) and temporal domain (percentile) were extracted. The Libsvm toolbox [41] with sigmoid kernel degree and the LDA classifier from Matlab were used. The LOOCV was applied in the classification performance assessment.

In our previous work, the minimum redundancy maximum relevance (mRMR) algorithm was used to select the most relevant features for multiclass sleep staging [21]. The same selected features and channels are used in this study. The forward/backward algorithm shows the contribution of each of these features and channels when combined with each other. In each decicion node of the decision tree, the combination of the most discriminative is selected. Table 5 presents the accuracy of the classification results. Table 6 shows the contribution of each channel in the classification, and shows the best

Table 6
Channels selection per decision tree node (classification accuracy using BDT-SVM, average results of all subjects).

Channels	Decision node			
	Wake vs. sleep	REM vs. NREM	N3 vs. N1+N2	N2 vs. N1
C3	86.54	90.41	90.24	87.81
C3C4	88.97	90.22	90.94	87.10
C3C4O1	91.73	93.73	91.08	87.78
C3C4O1O2	93.12	93.67	91.10	88.37
C3C4O1O2F3	94.39	93.89	91.85	88.48
C3C4O1O2F3F4	94.31	95.05	91.96	89.06
C3C4O1O2F3F4LOC	94.45	94.88	91.05	89.32
C3C4O1O2F3F4LOCROC	94.35	95.07	91.71	89.35
Selection:				
C3C4O1O2F3LOC	94.52	–	–	–
C3C4O1O2F3F4ROC	–	95.43	–	–
C3C4O2F4ROC	–	–	92.19	–
C3O1O2F3F4LOCROC	–	–	–	89.39

combination of channels for the different nodes. The best performance for classifying wake stage was obtained using the relative power features from all channels except F4 and ROC. To classify the REM stage, the best combination was found with relative power and harmonic parameters from all channels except LOC. The features used to classify N3 sleep stage were the relative power and the Harmonic parameters from all channels except O1, F3 and LOC. In the last node, all extracted features excluding the ones from channel C4 were used.

The set of selected features and channels presented in Tables 5 and 6 using BDT-SVM is coincident with the selection results using BDT-LDA classification. Therefore, the approach adopted for feature selection showed to be independent of the classification method. BDT-LDA and BDT-SVM classifiers were compared. The classification performance along the decision tree nodes was very similar, yet SVM classifier exhibited scores slightly higher, which motivated its use for the remaining analysis.

4.2. Post-processing: dubious epochs analysis

The post-processing step is almost immediate, since the application of the rules does not require intensive computation. The set of rules for the identification of dubious range were applied for every epoch classified by the BDT-SVM. The epochs that were selected as dubious were discarded and the remaining were classified again, thereby inferring the robustness of the dubious range detection (Table 7). The percentage of N1 epochs correctly assigned as N1 by the SVM classifier is less than 50%, and the percentage of epochs assigned as wake, N3 and REM sleep stages is just slightly lower than the percentage assigned by the expert. Epochs of these sleep stages are usually misclassified as N2. The DRD module could identify epochs falling in this situation. Applying the DRC the

differences are attenuated for all sleep stages, mainly for N2 (see H_{out} in Table 7).

STD-SVM misclassifications are mainly related with sleep transitions that are wrongly detected as going from REM to N2 instead of staying in REM, as inferred from Table 8. In our data, it is verified that the BDT-SVM classified around 22% of the epochs in REM as changing to N2 in the next epoch. However, in the scoring made by the expert, the percentage of epochs in this situation is almost inexistent. Table 8 complements the evidence of misclassification given in Table 7 showing the transitions between sleep stages using the different approaches to build the hypnogram; Table 8 shows not only the sleep stage with higher incidence of misclassifications but also between which stages these misclassifications occur for the 3 classification approaches used to build the hypnogram.

The percentage of sleep transitions from other stages to wake obtained by the SDT-SVM classification is similar to the expert scoring. The suggested correction improves the accuracy of all epochs except for those related to transitions from N1 to wake (in this case, the DRC rules introduce some errors). Regarding transitions to N1, the SVM and expert classifications differ mainly in epochs which are again N1 in subsequent epochs. The DRC improves all results except the misclassifications due to transitions from REM to N1. Using STD-SVM, a significant number of transitions occur from other sleep stages to N2, as can be seen by the higher percentages of transitions obtained with this classifier. Furthermore, the SDT-SVM considers that N2 epochs are changing to other sleep stages when, they actually do not. These values explain the situations highlighted in Table 8. DRD identifies most of the situations and DRC corrects mainly those related with transitions from wake, N1 and REM.

Table 9 describes the results of the overall approach per subject, presenting some detailed results that help to understand how the proposed system performance varies. The SDT-SVM accuracy for wake stage is higher than 90% for the most of the subjects (10 subjects). In fact, for these cases, the average classification accuracy for wake improved about 5% by applying the suggested rules. Considering only the epochs which were not selected as dubious, accuracy is on average about 97% (ranging from 94% to 99%). For subjects 8, 9, 11 and 14, the wake classification accuracy is relatively low. These subjects have multiple arousals during sleep, which may explain these results [42].

The lowest classification accuracy occurs for the N1 sleep stage. The DRC improved the accuracy results by 12% on average (ranging from 3% in subject 13 to 36% in subject 7). Discarding epochs in the dubious range, the classification accuracy rises 4%. Misclassifications of N1 were correctly identified in DRD, since the corresponding classification accuracy, discarding the dubious epochs, never decreased for any of the subjects. The low accuracies are related with the general low percentage of epochs on this stage to train the classifier, and also because the post-processing step can not detect all misclassifications.

Despite the positive overall assessment of the N2 classification post-correction, this step is also introducing some errors (take subject 7 as an example). The N2 corrections were especially efficient in subjects 4 and 9. However, the errors introduced in subjects 5, 6 and 7 show that the subjects' variability is still influencing negatively the global performance.

Table 7
Average percentages of sleep stages according to different classification approaches.

	Wake	N1	N2	N3	REM
H_{EM}	27.3	11.2	30.5	21.0	10.1
H_{in}	26.7	4.5	40.4	20.3	8.1
H_{nd}	26.1	4.2	29.3	18.6	7.4
H_{out}	26.9	9.9	31.2	20.8	11.1

Table 8
Average percentages of sleep stages transitions from stage i to stage j during one sleep night.

(*i*,*j*)%	**Wake**			**N1**			**N2**			**N3**			**REM**		
	H_{EM}	H_{in}	H_{out}	H_{EM}	H_{in}	H_{out}	H_{EM}	H_{in}	H_{out}	H_{EM}	H_{in}	H_{out}	H_{EM}	H_{in}	H_{out}
Wake	87	84	86	11	4	11	1	10	3	0	0	0	0	2	0
N1	15	15	20	46	30	39	35	47	36	0	0	0	4	8	5
N2	4	7	5	7	4	9	83	77	73	5	8	11	1	5	2
N3	1	1	1	1	0	1	6	14	13	92	84	84	0	0	0
REM	3	7	5	5	4	9	0	22	5	0	1	0	92	67	81

Table 9
Sleep classification accuracy results per subject (values rounded to units). ep – epochs not included in dubious range; pe – percentage of dubious epochs which are in fact TD-SVM misclassification; de – percentage of detected misclassifications; co – percentage of corrected misclassifications.

	Wake			**N1**			**N2**			**N3**			**REM**			**Total**			
	H_{in}	H_{out}	H_{nd}	H_{in}	H_{out}	H_{nd}	H_{in}	H_{out}	H_{nd}	H_{in}	H_{out}	H_{nd}	H_{in}	H_{out}	H_{nd}	ep	pe	de	co
1	96	97	98	52	70	54	82	87	87	92	94	94	72	94	92	86	93	56	43
2	95	95	97	67	73	70	86	87	89	91	96	97	90	96	96	91	68	36	16
3	91	92	96	59	73	68	83	87	90	89	91	91	84	94	97	80	85	61	31
4	90	88	94	60	72	65	79	87	90	90	94	95	86	94	97	76	62	63	36
5	97	98	98	50	63	50	83	78	82	83	86	86	81	87	86	95	73	27	11
6	94	95	96	70	79	78	89	84	92	96	97	97	83	94	97	87	55	43	0
7	98	98	98	50	86	50	92	81	94	97	98	99	100	100	100	95	60	63	0
8	74	77	83	59	69	63	76	80	82	87	89	90	92	93	96	80	69	41	17
9	82	79	84	55	59	58	74	84	84	88	92	93	90	94	96	75	72	56	26
10	92	94	99	50	70	50	77	83	88	80	89	89	81	91	92	75	73	60	29
11	79	81	86	57	69	61	73	78	80	76	79	79	73	89	87	80	76	41	19
12	94	95	96	67	74	70	86	86	87	57	60	58	72	80	78	94	69	20	12
13	94	94	94	63	66	65	86	84	88	87	94	93	100	100	100	93	44	15	0
14	69	69	72	71	74	77	66	72	74	91	92	93	79	93	92	81	73	42	17

With DRC the N3 classification accuracy has improved 3% on average. The largest increase in performance obtained with the DRC module with respect to N3 stage was reached for the subject 10 (about 9%). This may be explained due to the high percentage of errors detected in N3 (which is shown by the classification accuracy of epochs outside the dubious range). The lowest N3 classification accuracy occurred for subject 12, who also had the lowest REM classification accuracy. This is an unsual case of misclassification between N3 and REM. This subject has a very low percentage of REM sleep, and the automatic system is misclassifying N3 epochs as REM. Some of these errors are being detected and corrected in post-processing, but still in a low percentage.

The main part of misclassified REM epochs is well corrected. In 10 of the 14 subjects, REM stage is classified with an average accuracy of 95% (ranging from 92% to 100%) in epochs not included in the dubious range. The epochs are classified with an average accuracy of 96% (ranging from 91% to 100%) when DRC is applied. The low classification accuracy of the worst 3 cases (subjects 5, 11 and 12) is justified by the high incidence of misclassifications of REM transitions to other stages, as shown in Table 8. In some cases, the effect of DRC is null, which is related to the error introduced by the correction rules to N2 sleep stage and wake sleep stages, which annul the positive effect of the correction to the other stages.

On average, the overall misclassifications decreased 23.4% in 11 of the 14 subjects (ranging from 0% to 43%). The DRD detects 45% of the overall misclassifications (ranging from 15% to 63%). Furthermore, 70% of the epochs marked as dubious are in fact misclassified (ranging from 44% to 93%). Also, 85% of the epochs correctly classified are not included in the dubious range (ranging from 75% to 95%). Table 10 summarizes the overall classification accuracy. The classification performance of the epochs not selected for a dubious range is around 95% for wake and REM stages. With the inclusion of the DRC module, an improvement of about 10% in classification accuracy was obtained for N1 and REM sleep stages. With the inclusion of the DRC module, an improvement of about 10% in classification accuracy was obtained for N1 and REM sleep stages.

Table 10
Global accuracy results for the 14 subjects of the dataset.

	Wake	N1	N2	N3	REM	Total
H_{in}	92.74	60.43	81.15	87.70	83.16	81.03
H_{nd}	95.34	65.52	87.46	91.03	94.39	86.75
H_{out}	93.46	71.03	84.31	90.72	93.43	86.59

4.3. ASSC analysis tool

Fig. 6 exemplifies the sleep classification results considering the ASSC process without the post-processing (A), without considering the epochs selected as dubious (B) and after the suggested corrections to the dubious epochs (C). BDT-SVM results contain evident misclassifications between REM and N2, between N1 and N2 and between N2 and N3. The sleep hypnogram without dubious epochs (B) has an overall appearance close to that resulting from visual

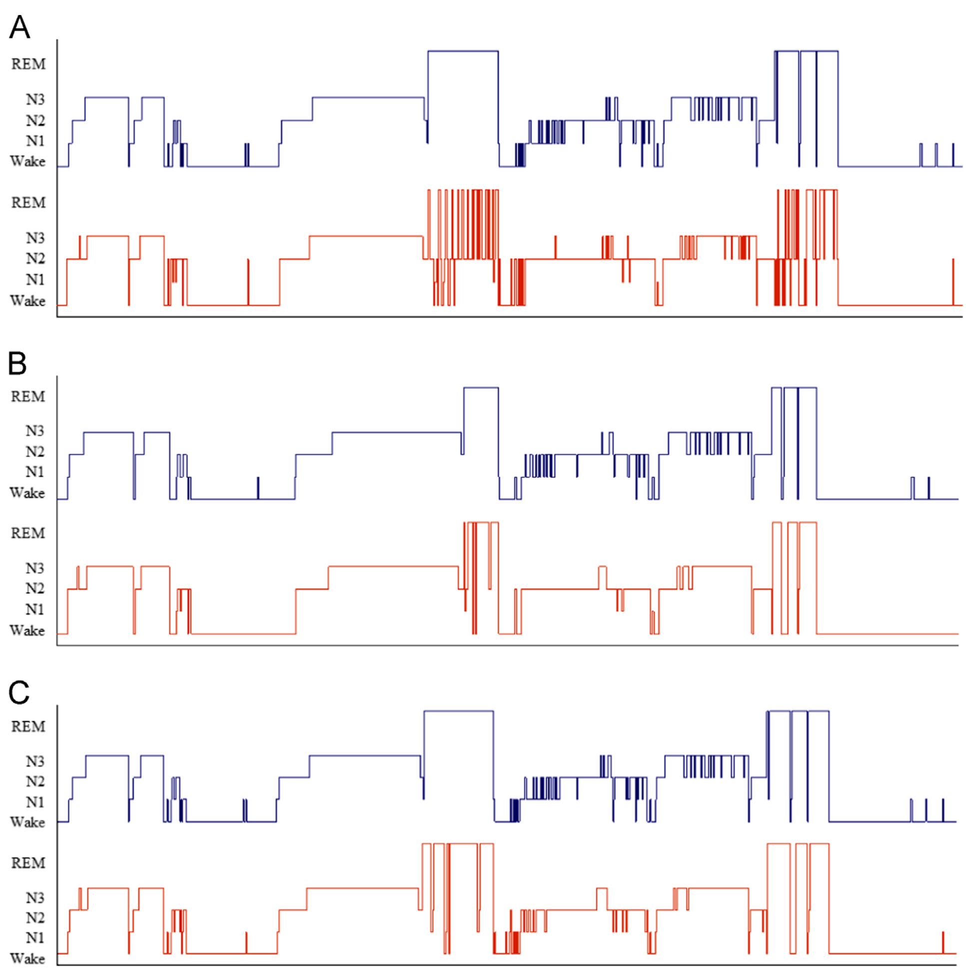

Fig. 6. Example illustrating the classification results. (A) Expert visual scoring (top) vs. BDT-SVM scoring (bottom). (B) Expert visual scoring (top) vs. BDT-SVM scoring without dubious range (bottom). The graph shows only the epochs that are not included in dubious range (86% of the total number of epochs). (C) Expert visual scoring (top) vs. BDT-SVM scoring after post-processing (bottom).

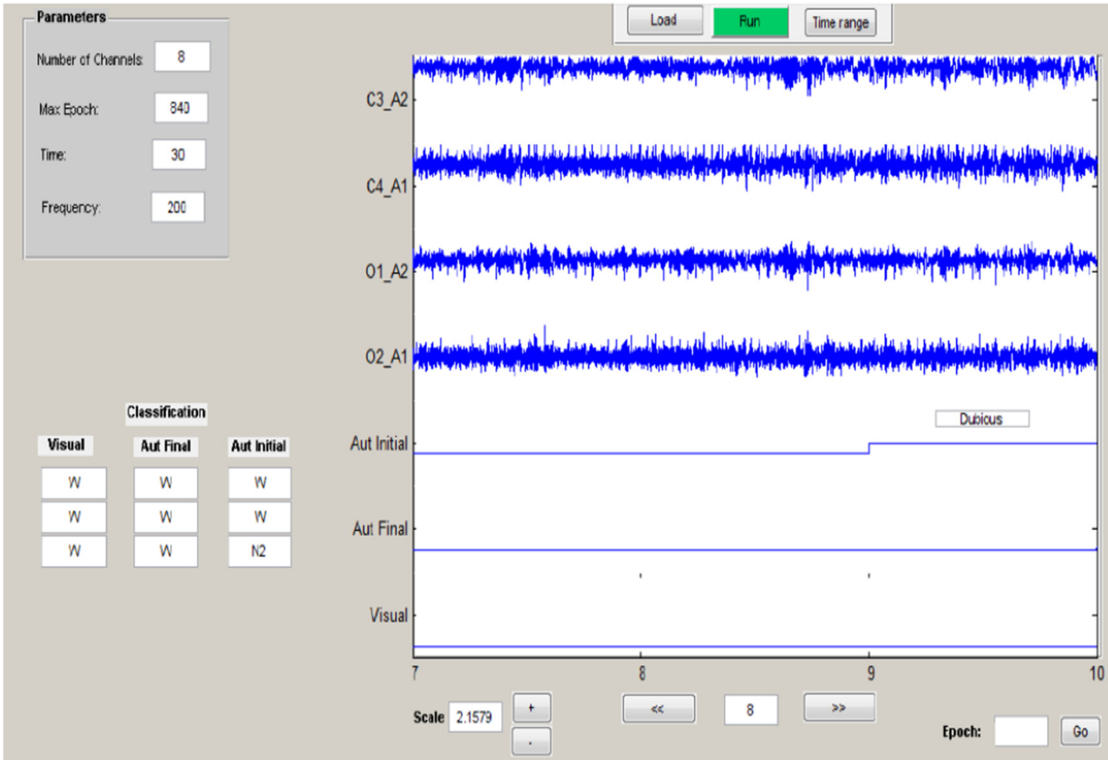

Fig. 7. Illustration of the ASSC analysis tool. Aut Initial (H_{in}) – SVM classification is shown as hypnogram with epochs with high probability of misclassification marked as dubious. Aut Final (H_{out}) – automatic classification applying DRD and DRC. Visual (H_{EM}): expert classification. The classification labels are shown in the table on the left of the graph. The parameters related to the EEG and EOG channels: number of channels and epochs available to display, duration of each epoch and sampling frequency.

expert scoring. However, still contains some misclassifications, mainly between N2 and N3. After DRC (C), misclassifications related with sleep transitions were almost eliminated. The remaining errors are due to non-detection by DRD (see Fig. 6B).

Fig. 7 shows a snapshot of our ASSC analysis tool. It shows the hypnogram obtained from BDT-SVM classification (Aut Initial (H_{in})), tagging dubious epochs. The resulting hypnogram, after DRC has been applied, is also shown (Aut Final (H_{out})). The visual scoring made by the expert is shown only for comparison. In the example, for 3 epochs, the first 2 are correctly classified and none is marked as dubious. In the third epoch the subject is awake and the SVM classified it as N2. The post-processing step tags this epoch as dubious and suggests that the correct classification is wake.

5. Discussion

In this work, we take advantage of the knowledge based on standard sleep patterns and AASM rules to propose an ASSC approach with two classification outputs: the non-dubious epochs (around 80% of the data with low probability of misclassification) and the dubious epochs (epochs suspected of misclassification for which it is suggested a second label of classification).

5.1. Feature and classifier selection

The first step of the proposed algorithm is based in the standard patterns recognition algorithms. The classification step is performed according to a binary decision tree structure, suitable to find the most discriminative features, depending on the sleep stage to classify. The combination of features and channels per decision node varies. For example, in wake vs. sleep classification, the best performance in the classification was attained using only relative power features, but for REM vs. NREM classification the Harmonic parameters are also discriminative, and for N1 vs. N2 classification the MODWT provides relevant information.

In all decision nodes, the relative power features from at least one EEG channel per brain lobe and one EOG channel are relevant. The results also showed that there is a redundancy in using both EOG channels. To classify the wake stage, the relative power is enough, and the less relevant channels are the frontal EEG and EOG. The most important features come from alpha activity in occipital and central channels (Table 4). In sleep stages N1 and N2, all extracted features are relevant except those from the central channels (one channel is enough). These stages present a wide variety of patterns [5], which hinders their correct classification. According to the results, the relative power and harmonic parameters of central channels provide the most relevant information for the detection of stage N3. From EOG, occipital and frontal channels, it is enough to use only one channel, which is explained by the fact that deep sleep is characterized by slow waves at all brain (Tables 1 and 3). The REM stage is classified based on relative power and on harmonic parameters of all channels except LOC. It is a complex stage requiring information from almost all channels given that it is characterized by diverse patterns spatially distributed; the highest regular patterns are related with ocular activity [5]. The attained conclusions from experimental results, in respect to the most important channels and features are in agreement with the AASM rules (Table 1) [5].

The processing time for the classification of a complete overnight record was less than 2 min for both LDA and SVM classifiers. Nevertheless, SVM performed globally better than LDA, and thus it was used to obtain the best classification results.

5.2. Post-processing: dubious range analysis

As hypothesized, the post-processing based on the standard sleep classification rules led to a classification improvement of the first step of automatic classification based on SVM or LDA. Our results suggest

that the inclusion of the knowledge based on standard AASM classification rules helps to automatically identify and solve some common problems of the ASSC systems, mainly misclassifications between N2 and REM that occur due to the high number of transitions between those stages.

The rules of the DRD and DRC were designed from empirical and analytical models, based on a prior analysis of typical errors in ASSC. Not all AASM rules were useful to apply on this approach, neither the used rules were always efficient (sometimes the DRC effect is null). One plausible explanation is that the features extracted from epochs with duration of 30 s do not give enough information about signal evolution that can occur within a singular epoch.

The misclassification problems are clearly related with epochs in N1, N3 and REM classified as N2 (Table 7). These stages present common neurophysiologic patterns (Table 3), and are all subsequent stages of N2 in sleep evolution (Fig. 4). Analyzing also the sleep transitions in Table 8, it was verified that the high percentage of BDT-SVM misclassifications is due to REM epochs classified as N2.

The majority of detected misclassifications of N3 and REM are corrected by post-processing. The sleep misclassifications easiest to detect are those related with REM and N2. However, the highest correction percentage is for N1 and REM (Table 10). Thus, the optimization of the proposed algorithm is mainly dependent of the optimization of the post-processing step to N2 sleep stage.

N1 and N2 are the sleep stages which present more differences between clinical expert and automatic classifications. Therefore, the overall classification performance to these stages is lower than for the other cases. N1 sleep stage misclassifications assume a higher influence in relation to the other sleep stages misclassifications due to the low number of epochs in this stage per subject.

Comparing with the state-of-the-art, the proposed ASSC system shows as a main advantages: the highest accuracy to classify wake and REM sleep stages of epochs not marked as dubious; the detection of 45% of the total misclassifications; and the ability to correct 23.4% of them. Rigorous direct comparisons between classification performance of the proposed algorithm and the algorithms of the works cited in Table 2 could be misleading. However, the proposed classification process presents globally a performance with accuracy levels similar or higher than the state-of-the-art.

6. Conclusion

In this work, a novel two-step ASSC approach was proposed, aiming to overcome problems related to ambiguous activity and subjects' variability. The algorithm is oriented to a post-processing step based on error analysis according to standard sleep classification rules.

By identifying a dubious range of classification, the automatic classification of wake and REM stages reaches high levels of accuracy (around those required for clinical applications). A suggestion of correction is given to the epochs marked as dubious, thereby improving the average results of all sleep stages. Despite the general improvements, some classification problems persist, mainly misclassifications of sleep stage transitions. The classification performance levels required to apply in clinical practice were achieved in some sleep stages for several subjects.

This approach provides the expert a reliable ASSC for non-dubious epochs (which are most of the PSG records). Further research and optimization in this direction, with the incorporation of additional domain-knowledge using more sophisticated expert-based rules and new parameters of patient profile definition, may well lead to increase performance up, mainly in distinction between N1 and N2 sleep stages, to a satisfactory level of applicability on heterogeneous EEG and EOG records.

The proposed system has been developed and assessed using data collected from patients with suspected sleep disorders. To apply in patients already diagnosed and under treatment, it will be necessary an optimization of the proposed system. As each disorder and treatment can be associated with unusual sleep patterns, we intend as future work to embed in our algorithm, knowledge about specificities of several sleep disorders.

Summary

The main challenging issues of automatic sleep staging have been the subject's variability and the common sleep patterns between different sleep stages.

In this paper, we suggest a system which provides two outputs: epochs with high probability of correct classification (non-dubious classification) and epochs with high probability of misclassification (dubious classification). In the automatic sleep hypnogram, the epochs with dubious classification are marked, and then they are relabeled according to a post-processing correction algorithm. With this approach, the expert physician receives an hypnogram that indicates the epochs classified with a high degree of certainty and the epochs with ambiguous classification (epochs for which a final manual scoring cycle is required).

The system is divided into two main parts: one first step of classification, based on a typical process performed by an automatic pattern recognition algorithm, and a second step of post-processing classification based on heuristic rules, which incorporates information from the standard rules of sleep classification, aiming to detect and correct misclassifications. Six electroenceplahographic channels (F3, C3, O1, F4, C4, O2) and two electroocolugraphic channels, right and left (ROC, LOC), are used to classify 5 stages (Wake, NREM sleep – N1, N2 and N3, and REM sleep).

Our goal is to provide a solution as generic as possible that can be useful in sleep staging of patients with suspected sleep disorders. The proposed system was tested in a dataset of 14 clinical records. Wake and REM epochs not falling in the dubious range are classified with accuracies in levels required for clinical applications. The post-processing score reassignment of epochs tagged as dubious enhances the global results of all sleep stages. However, some typical problems of the automatic sleep stage classification, as it is the case of the low number of N1 epochs to train the models, continue affecting the classification performance. Another critical issue is related to the variations on the sleep patterns regarding the different sleep disorders and treatments. We intend, as future work, to prepare our algorithm to incorporate knowledge about the specificities of several sleep disorders and test it in a larger dataset.

Conflict of interest statement

None declared.

Acknowledgments

This work was supported by the Portuguese Foundation for Science and Technology (FCT) under Ph.D. Grants SFRH/BD/80735/2011, SFRH/BD/81828/2011, and FCT project AMS-HMI12: RECI/EEIAUT/0181/2012, cofounded by COMPETE. We also thank the Sleep Medicine Centre of Coimbra University Hospital Centre (CHUC) the provided data. We are also very grateful to José Moutinho dos Santos and Mafalda Ferreira for the help in PSG records understanding.

References

[1] A. Roebuck, V. Monasterio, E. Gederi, M. Osipov, J. Behar, A. Malhotra, T. Penzel, G.D. Clifford, A review of signals used in sleep analysis, Physiol. Meas. 35 (1) (2014) R1–57.
[2] J. Santamaria, B. Högl, C. Trenkwalder, D. Bliwise, Scoring sleep in neurological patients: the need for specific considerations, Sleep 34 (10) (2011) 1283–1284.
[4] F. Chapotot, G. Becq, Automated sleep–wake staging combining robust feature extraction, artificial neural network classification, and flexible decision rules, Int. J. Adapt. Control Signal Process. 24 (2010) 409–423.
[5] American Academy of Sleep Medicine, AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, American Academy of Sleep Medicine, Westchester, IL, 2007.
[6] M. Silber, S. Ancoli-Israel, M.H. Bonnet, et al., The visual scoring of sleep in adults, J. Clin. Sleep Med. 3 (2) (2007) 121–131.
[7] R. Agarwal, J. Gotman, Computer-assisted sleep staging, IEEE Trans. Biomed. Eng. 48 (2001) 1412–1423.
[8] L. Zoubek, S. Charbonnier, S. Lesecq, A. Buguet, F. Chapotot, Feature selection for sleep/wake stages classification using data driven methods, Biomed. Signal Process. Control 2 (2007) 171–179.
[9] L. Fraiwan, N. Khaswaneh, K. Lweesy, Automatic sleep stage scoring with wavelet packets based on single EEG recording, Proc. World Acad. Sci., Eng. Technol. 54 (2009) 485–488.
[10] V. Helland, A. Gapelyuk, A. Suhrbier, M. Riedl, T. Penzel, J. Kurths, N. Wessel, Investigation of an automatic sleep stage classification by means of multi-scorer hypnogram, Methods Inf. Med. 4 (2010) 1–6.
[11] M. Ronzhina, O. Janoušek, J. Kolářová, M. Nováková, P. Honzík, I. Provazník, Sleep scoring using artificial neural networks, Sleep Med. Rev. 16 (2012) 251–263.
[12] G. Han, J. Park, C. Lee, et al., Genetic fuzzy classifier for sleep stage identification, Comput. Biol. Med. 40 (2010) 629–634.
[13] S. Günes, K. Polat, S. Yosunkaya, Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting, Expert Syst. Appl. 37 (2010) 7922–7928.
[14] L. Doroshenkov, V. Konyshev, S. Selishchev, Classification of human sleep stages based on EEG processing using hidden Markov models, Biomed. Eng.41 (2007) 25–28.
[15] J. Dong, D. Liu, Automated sleep staging technique based on the empirical mode decomposition algorithm: a preliminary study, Adv. Adapt. Data Anal. 2 (2010) 267–276.
[16] G. Garg, V. Singh, M. Grover, J. Gupta, Optimal Kernel learning for EEG based sleep scoring system, Int. J. Biol. Med. Res. 2 (2011) 1220–1225.
17 B. Koley, D. Dey, An ensemble system for automatic sleep stage classification using single channel EEG signal, Comput. Biol. Med.42 (2012) 1186–1195.
[18] S. Charbonnier, L. Zoubek, S. Lesecq, F. Chapotot, Self-evaluated automatic classifier as a decision-support tool for sleep/wake staging, Comput. Biol. Med. 41 (6) (2011) 380–389.
[19] T. Sousa, D. Oliveira, S. Khalighi, G. Pires, U. Nunes, Neurophysiologic and statistical analysis of failures in automatic sleep stage classification, in: Proceedings of the BIOSIGNALS – International Conference on Bio-inspired Systems and Signal Processing, 2012, pp. 423–428.
[20] Y. Hsu, Y. Yang, J. Wang, C. Hsu, Automatic sleep stage recurrent neural classifier using energy features of EEG signals, Neurocomputing 104 (2013) 105–114.
[21] S. Khalighi, T. Sousa, D. Oliveira, G. Pires, U. Nunes, Efficient feature selection for sleep staging based on maximal overlap discrete wavelet transform and SVM, in: Proceedings of the 33rd Annual International IEEE EMBS Conference (EMBC11), 2011.
[22] A. Krakovská, K. Mezeiová, Automatic sleep scoring: a search for an optimal combination of measures, Artif. Intell. Med.53 (2011) 25–33.
[23] V. Bajaj, R. Pachori, Automatic classification of sleep stages based on the time-frequency image of EEG signals, Comput. Methods Programs Biomed. 112 (2013) 320–328.
[24] S. Khalighi, T. Sousa, U. Nunes, Adaptive sleep stage classification under covariate shift, in: Proceedings of the IEEE 34th Annual International EMBS Conference (EMBC 2012), 2012.
[25] S. Liang, C. Kuo, Y. Hu, Y. Cheng, A rule-based automatic sleep staging method, J. Neurosci. Methods 205 (2012) 169–176.
[26] S. Khalighi, T. Sousa, G. Pires, U. Nunes, Automatic sleep staging: a computer assisted approach for optimal combination of features and polysomnographic channels, Expert Syst. Appl. 40 (2013) 7046–7059.
[27] American Academy of Sleep Medicine, International Classification of Sleep Disorders, Revised: Diagnostic and Coding ManualAmerican Academy of Sleep Medicine (2001) .
[28] Jaakko MalmivuoRobert Plonsey, Bioelectromagnetism – Principles and Applications of Bioelectric and Biomagnetic Fields, Oxford University Press, New York, 1995.
[29] J. Muthuswamy, N. Thakor, Spectral analysis method for neurological signals, J. Neurosci. Methods 83 (1998) 1–14.
[30] B.D. Percival, T.A. Walden, Wavelet Methods for Time Series Analysis, Cambridge University Press, Cambridge, 2000.
[31] F. Mormann, R.G. Andrzejak, C.E. Elger, K. Lehnertz, Seizure prediction: the long and winding road, Brain 130 (2007) 314–333.
[32] W.C. Tang, S.W. Lu, C.M. Tsai, C.Y. Kao, H.H. Lee, Harmonic parameters with HHT and wavelet transform for automatic sleep stages scoring, Proc. World Acad. Sci., Eng. Technol. (2007) 414–417.
[33] G. Becq, S. Charbonnier, F. Chapotot, A. Buguet, L. Bourdon, P. Baconnier, Comparison between five classifiers for automatic scoring of human sleep recordings, in: S. K. Halgamuge, L. Wang (Eds.), Studies in Computational Intelligence (SCI), Classification and Clustering for Knowledge Discovery 2005, pp. 113–127.
[34] S. Aksoy, R. Haralick, Feature normalization and likelihood based similarity measures for image retrieval, Pattern Recognit. Lett. 22 (2001) 563–582.
[35] Richard O. Duda, Peter E. Hart, David G. Stork, Pattern Classification, 2nd Edition, Jonh Wiley and Sons LTd., 2000.
[36] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods), Cambridge University Press, Cambridge, 2000.
[37] M. Kryger, T. Roth, W.C. Dement, Principles and Practice of Sleep Medicine, Elsevier/Saunders, Philadelphia, 2005.
[38] D. Mazzotti, C. Guindalini, W.A. Moraes, M.L. Andersen, M.S. Cendoroglo, L.R. Ramos, S. Tufik, Human longevity is associated with regular sleep patterns, maintenance of slow wave sleep, and favorable lipid profile, Front. Aging Neurosci. 6 (134) (2014) 1–9.
[39] B. Clarke, E. Fokoue, H.H. Zhang, Principles and Theory for Data Mining and Machine Learning, in Springer Series in Statistics, S.S.B. Media, 2009.
[40] K.H. Brodersen, C.S. Ong, K.E. Stephan, J.M. Buhmann, The balanced accuracy and its posterior distribution, in: Proceedings of the Pattern Recognition (ICPR), 2010, pp. 3121–3124.
[41] C. Chang, C. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (2011) 1–39.
[42] M.H. Asyali, R. Berry, M. Khoo, A. Altinok, Determining a continuous marker for sleep depth, Comput. Biol. Med. 37 (11) (2007) 1600–1609.